Inference Control in Medicine

Inference control in medicine

- Background: medical data is often used for research purposes
- Method: data is anonymised by removing personal data (names, addresses)

Inference control (2)

- However, queries can be made via secondary fields
  - Dates
    - Birthdate
    - Date of treatment (e.g. known accident)
  - Places
  - Known diseases
  - Relationships
    - Family

Inference control in census info

- Census collects much sensitive information
- (However, it has fewer fields and usually different queries than medical data)

De-identify before processing

- Make available one record in thousand
  - Minus names, addresses, etc.
- Add noise
  - Prevent identifying individuals
- Supress records with extreme values
  - Also protects individuals

Restrict available queries

- Database contains full data set
- Security comes from restricting the kinds of queries that can be performed
- However, there are several attacks
  - Make queries to samples containing given person
  - Infer from statistics particular information about that person
  - Such queries are called *trackers*

Related problem: aggregation

- Combining (potentially large number of) unclassified information to obtain classified information
- Example: undercover operatives can be inferred from phone book and postings
- Aggregation of several databases can make trackers easier

Inference control mechanisms

Minimum query size

- Do not answer queries which contain fewer than *n* records
- However, attack:
  - Make query to *n* records
  - Make query to *n* records + target
- Also: restrict queries returning almost all records
  - Answer queries where $n < count < N-n$

Trackers

- Individual tracker: based on some specific characteristic of target individual
  - e.g. only female, only one with PhD etc.
- General tracker: can find out any statistic
  - Minimum query set is *n*
  - Number of statistics is *N*
  - If $n < N/4$ and no restrictions on queries, we can create general trackers

More query controls

- *n*-respondent, *k%*-dominance rule
  - do not show statistic where *k%* or more is contributed by *n* or fewer values
- Suppress data with extreme values
- Suppress more sensitive statistics on local level
  - However, for global average, these can be included

Cell suppression

Major / Minor	**Biology**	**Physics**	**Chemistry**	**Geology**
Biology	-	16	17	11
Physics	7	-	32	18
Chemistry	33	41	-	2
Geology	9	13	6	-

Cell suppression

Major / Minor	**Biology**	**Physics**	**Chemistry**	**Geology**
Biology	-	16	17	11
Physics	7	-	32	18
Chemistry	33	41	-	2
Geology	9	13	6	-

Cell suppression

Major	**Biology**		**Physics**		**Chemistry**		**Geology**	
Minor								
Biology	-		16		17		11	
Physics	7		-		32		18	
Chemistry	33		41		-		2	
Geology	9		13		6		-	

Cell suppression

Major	**Biology**		**Physics**		**Chemistry**		**Geology**	
Minor								
Biology	-		16		17		11	
Physics	7		-		32		18	
Chemistry	33		41		-		2	
Geology	9		13		6		-	

Cell suppression

- If database contains *m*-tuples, then blanking single cell generally means suppressing 2^m-1 other cells
- For online queries, use implied queries control
  - Query on *m* attribute values is allowed if all of the possible query sets have at least *k* records

Maximum order control

- Limit number of attributes that a query can have
- Example: 1000 medical records
  - 3 attributes were safe
  - 4 attributes: one individual record
  - 10 attributes: most records

Query overlap control

- Keep track of previous queries
- Reject query from user if results could be combined with previous results could disclose sensitive information
- Problems
  - Computationally expensive
  - Users can cooperate

Randomization

- Perturbation – add noise with zero mean and known variance
- Random sample queries – all the responses are the same size
  - Released data is computed from randomly selected subset of all records

Active attacks

- So far, we have discussed *passive* attacks that do not modify database
- *Active* attacks modify database
  - Add records to make group containing target record and known records
- Active attacks can also modify environment
  - Set drug prices so that use patterns can be determined from statistics

Payment information

- Payment information is usually personalized
- Treatments and prescribed medicine directly point to diagnosis
- Medical insurers often collect large amount of sensitive data
  - Insurer can be private organization
- Usual solution is to use legal regulation

Summary

- In medicine and intelligence, the real-time access control is manageable
- Hard problem is to prevent inference of sensitive data
- Also difficult is supporting research based on statistics

Summary (2)

- Much depends on attack model
  - Detailed information about specific person?
  - Detailed information about any person?
  - Less-detailed information about large amount of people?
- Centralization of resources has effect on potential losses
  - Decentralized: smaller losses
  - Centralized: higher losses (but less probable)